DOCUMENT RESUME

ED 076 604 TM 002 584

AUTHOR	Oosterhof, Albert C.
TITLE	Stability of Various Item Discrimination Indices.
PUB DATE	1 Mar 73
NOTE	16p.; Paper presented at Annual Meeting of the American Educational Research Association (New Orleans, Louisiana, February 25 - March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Discriminant Analysis; Grade 10; *Item Analysis; *Psychometrics; Technical Reports; *Test Construction; Test Reliability

ABSTRACT

The stability of various item discrimination indices when a particular item assumes membership within different sets of items was investigated. Results suggest that when item discrimination is used as a significant factor in determing which items from several forms will be selected for inclusion in a test, the actual items selected are less a function of the chance combinations of items which made up the various forms when Gulliksen's item reliability index and Findley's difference index are used as indices, than when phi coefficients, and particularly point biserial and biserial coefficients are used as estimates of item discrimination. (Author)

FORM 850
PRINTED IN U.S.A.

FILMED FROM BEST AVAILABLE COPY

ED 076604

TM 002 584

Stability of Various Item Discrimination Indices[1]

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Albert C. Oosterhof

Florida State University

The development of a test consists of several individual steps. After the domain to be measured has been defined and a large number of test items have been written, the next step involves the collection of data as a basis for selecting the best available items to form the final test. Henrysson (1971) recommends three tryout administrations of items in the development of the test, the first involving the use of a small group of examinees to discover gross defects in items, the second for determining relevant item parameters to the pool of items in order to select items to be included in the final form of the test, and the third administration for discovering minor defects in the total instrument and determining test reliability.

The present study was concerned with the second of these administrations, and more specifically with the degree to which various discrimination indices remain invariant when the item with which they are associated assumes membership within different sets of items. The normal procedure used in this second tryout of items is to divide the total sample of items into several test forms and administer each form to a different group of subjects. One of the inherent assumptions involved in this procedure is that, as an item assumes membership within varying sets of items, the respective parameters associated with the item remain constant. If this assumption of stability is not true for a particular index of discrimination, then the items selected for inclusion in the final test form on the basis of such an index will be determined by the chance combination of items with which the item was originally associated, and not necessarily by the discriminatory power intrinsic to the item were it to be included in the final form of the test instrument.

Method

The item discrimination indices compared in this study include point biserial coefficients, biserial coefficients, phi coefficients, estimates of tetrachoric coefficients, Gulliksen's item reliability index, Findley's difference index, Flanagan's approximation and Davis' transformation of the product-moment correlation coefficient. An adaptation of the

---

[1]A paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, February 25 - March 1, 1973.

point-biserial coefficient using Livingston's (1972) criterion-referenced applications of classical test theory was obtained by substituting various criterion scores for the test mean. The various indices and forms of these discrimination indices included in the present study, the notation used to represent these indices, and appropriate references to the derivation of respective indices are listed in Table 1.

Source of Data

The Verbal Reasoning (VR) subtest of Form M to the Differential Aptitude Tests (DAT), 1963 Revision, was used as the source of items for this study. This particular subtest consists of 50 items and was administered to tenth-grade students in 11 school districts in the State of Kansas. The school districts participating in the present study were consolidated districts, each district enrolling students from a small city or town environment in addition to students from neighboring rural areas.

Procedure

From the total population of 2,311 tenth-grade students enrolled in the indicated districts, a sample of 2,000 subjects (Ss) were selected for inclusion in the study. These 2,000 Ss were grouped according to school districts, and then divided into two groups of 1,000 Ss each by taking sequential pairs of Ss, using a random number generator to assign the first member of each pair to one of the groups, and then assigning the remaining member to the other group. The first of these two groups was used for establishing estimates of the population parameters associated with each of the discrimination indices investigated in this study. The second group of 1,000 Ss was used for calculating the various discrimination indices for each item after the particular item had been assigned membership within varying sets of other items.

As indicated above, the first group of 1,000 Ss was used for establishing estimates of the population parameters associated with each of the various discrimination indices. Using the second group of 1,000 Ss, the discrimination indices were calculated for each item after the item had been assigned membership with varying numbers of other items. For example, the first of the 50 VR items was placed with nine items randomly selected from the remaining 49 VR items. This completed set of ten VR items constituted a total test and each of the various item indices listed above were computed for the first item. This procedure was repeated for each of the VR items, i.e., each item was placed in a set with nine other randomly selected items and the various item indices were computed for the originally selected item in the particular set of ten items.

Table 1

Listing of Discrimination Indices Investigated

Notation	Description of Index	Reference
	Point-Biserial Correlation Coefficient	
r_{pb}	Unadjusted coefficient	
r'_{pb}	Coefficient adjusted for spuriousness	Henrysson (1963)
r''_{pb}	Coefficient adjusted for spuriousness and different sums	Henrysson (1963)
S	Coefficient of selective efficiency (point-biserial coefficient divided by maximum obtainable point-biserial coefficient)	Brogden (1949)
	Biserial Correlation Coefficient	
r_b	Unadjusted coefficient	
r'_b	Coefficient adjusted for spuriousness	Henrysson (1963)
r''_b	Coefficient adjusted for spuriousness and different sums	
r_{Nb}	Coefficient calculated with continuous variable normalized by use of stanines	Henrysson (1971)
	Phi Correlation Coefficient with total scores dichotomized using subjects scoring in the	
$\emptyset_{.10}$	Upper and lower 10%	
$\emptyset_{.25}$	Upper and lower 25%	
$\emptyset_{.33}$	Upper and lower 33%	
$\emptyset_{.50}$	Upper and lower 50%	

Table 1

Listing of Discrimination Indices Investigated (continued)

Notation	Description of Index	Reference
	Tetrachoric correlation coefficient using subjects scoring in the	
$r_{.10}$	Upper and lower 10%	
$r_{.25}$	Upper and lower 25%	
$r_{.33}$	Upper and lower 33%	
$r_{.50}$	Upper and lower 50%	
$r_{it}s_i$	Gulliksen's Item Reliability Index	Gulliksen (1950)
D	Findley's Difference Index	Findley (1956)
r_F	Flanagan's approximation of the Product-Momert Correlation Coefficient	Flanagan (1939)
r_D	Davis' transformation of the Product-Moment Correlation Coefficient	Davis (1946)
	Point Biserial Correlation Coefficient Adjusted for use with Criterion-Tests with criterion score set at	Costerhof (1972)
$r_{pb,c=.25}$	25% mastery	
$r_{pb,c=.50}$	50% mastery	
$r_{pb,c=.75}$	75% mastery	
$r_{pb,c=.90}$	90% mastery	

The procedure outlined to this point provided information on the stability of the various item indices when a particular item is included in a set of ten items. In practice, one initially starts with some finite set of items from which the items making up the final form of the test are selected. Nunnally (1967) suggested that one should begin with at least twice the number of items as will appear in the final form of the test. Henryssen (1971), however, stated that this margin of discard could at times be as small as 20%. In order to determine the stability of the various discrimination indices included in the present study, the procedure discussed above was repeated using sets of 25 and sets of 40 items. In order to increase the confidence with which each index's invariance property could be discussed, the entire procedure was replicated for each item three times using different random sets of 10, 25 and 40 items on each occasion. All indices were calculated for each item, each index being based on the responses of the 1,000 Ss included in the second group.

Data Analysis

Each item within the VR subtest was assigned a rank position on the basis of the magnitude of the discrimination index associated with the respective item. Since there were a total of 24 indices associated with each item, this ranking was determined 24 separate times, once for each index. Again, since each of the 24 indices were calculated using sets of 10, 25, and 40 items, and each of these conditions were replicated three times with each set size, a total of 3 by 3 by 24 or 216 separate rankings of items were involved.

The rank-order correlation was computed between the population estimates of each discrimination index and the corresponding index which was calculated when each item was given membership within a set of items of a given number (i.e., 10, 25, or 40 items). The resulting rho coefficient indicated the degree of relationship between the population estimates of each index, and the corresponding values derived for each index when items were assigned membership within different sets of items. Therefore, the resulting rho coefficients indicated the degree of invariance associated with each of the 24 discrimination indices.

These 24 rho coefficients associated with each of the specified conditions were ranked in order to determine the relative stability of the 24 discrimination indices. As a result, three ranks were associated with each index, these ranks indicating the index's relative invariance for each of three replications which were executed using a particular set size. The median of three ranks associated with each index therefore indicate an index's relative invariance. Each index was associated with a total of three median ranks since indices were calculated with set sizes of 10, 25, and 40 items.

Results

The rank-order correlations between the estimated population values of each index and the corresponding values of each index obtained using sets of 10 items are listed in Table 2. The 24 discrimination indices were ranked according to their invariance as indicated by the rho coefficients. This ranking was done for each of the three replications made using random sets of 10 items, the least variant index within each replication being ranked 1. The median of the three ranks associated with each index was the basis of ordering the 24 indices in Table 2 from most stable to least stable. Tables 3 and 4 contain corresponding information for replications involving random sets of 25 and 40 items, respectively.

Gulliksen's item reliability index ($r_{it}s_i$) and Findley's difference index (D) were the most stable of the 24 indices with respect to the invariance characteristics investigated in the present study. The phi coefficients (Ø) calculated using Ss who scored in the upper and lower 27% or upper and lower 33% were ranked next below $r_{it}s_i$ and D with respect to invariance, however Ø coefficients appeared less stable when Ss' performances were dichotomized into groups representing the upper and lower 10%, and particularly less stable when dichotomized into groups representing the upper and lower 50%.

The point-biserial correlation coefficient (r_{pb}), when used as a discrimination index, generally ranked after $Ø_{.33}$ and $Ø_{.27}$ relative to its invariance characteristics. The various adjustments to the point-biserial coefficient displayed values of rho relatively close to that of the point-biserial before adjustment, however, almost without exception, the rho's associated with the unadjusted coefficient were the highest. As a group, the various forms of the biserial coefficient were more variant than the various forms of the point-biserial coefficient, however the difference between these two groups was relatively small.

The rho coefficients associated with the estimates of the tetrachoric correlations displayed a wide range of values, and with the exception of when computations were based on Ss scoring in the upper and lower 10%, the stability of the tetrachoric coefficient relative to other indices was quite unpredictable. Each replication typically ranked in the lower half of the indices relative to invariance, however some individual replications ranked as high as fourth and fifth with respect to stability. The estimates of the tetrachoric coefficients which were based on Ss scoring in the upper and lower 10% without exception demonstrated this particular form of r_t to be the least stable of the 24 indices.

TABLE 2

Rank-Order Correlations Between Population Estimates of the Various Discrimination Indices and Calculated Values of Corresponding Indices Obtained Using Sets of 10 Items

Index	Rho Coefficients (1)	(2)	(3)	Ranks of Rho Coefficients (1)	(2)	(3)	Median Rank
$r_{it}s_i$	.896	.907	.903	1.0	1.0	1.0	1.0
D	.824	.858	.883	5.5	2.0	2.0	2.0
$\phi_{.33}$	.868	.835	.862	2.0	3.0	3.0	3.0
r_{pb}	.840	.820	.818	3.0	4.5	5.0	4.5
r''_{pb}	.783	.820	.813	11.5	4.5	6.0	6.0
$\phi_{.27}$	.765	.805	.840	15.0	8.0	4.0	8.0
$\phi_{.50}$	.793	.816	.793	9.0	6.0	8.5	8.5
S	.824	.750	.793	5.5	14.0	8.5	8.5
$r_{pb,c=.50}$	.812	.735	.787	7.0	15.0	10.0	10.0
$r_{.50t}$	.833	.774	.722	4.0	10.5	17.0	10.5
r'_{pb}	.779	.811	.786	13.0	7.0	11.0	11.0
r_b	.783	.781	.720	11.5	9.0	18.5	11.5
$\phi_{.10}$	.802	.590	.744	8.0	23.0	12.0	12.0
r_{nb}	.791	.766	.747	10.0	13.0	13.0	13.0
r''_b	.767	.774	.744	14.0	10.5	14.0	14.0
r'_b	.764	.768	.720	16.0	12.0	18.5	16.0
$r_{.33t}$	.732	.648	.800	17.0	20.0	7.0	17.0
$r_{pb,c=.90}$	.707	.675	.664	18.0	17.0	22.0	18.0
r_D	.674	.663	.728	19.5	18.5	15.0	18.5
r_F	.674	.663	.727	19.5	18.5	16.0	18.5
$r_{.27t}$	.671	.714	.705	21.0	16.0	20.0	20.0
$r_{pb,c=.75}$	.668	.644	.634	22.0	21.0	23.0	22.0
$r_{pb,c=.25}$	.584	.619	.672	23.0	22.0	21.0	22.0
$r_{.10t}$	.479	.420	.563	24.0	24.0	24.0	24.0

TABLE 3

Rank-Order Correlations Between Population Estimates of the Various Discrimination Indices and Calculated Values of Corresponding Indices Obtained Using Sets of 25 Items

Index	Rho Coefficients (1)	(2)	(3)	Ranks of Rho Coefficients (1)	(2)	(3)	Median Rank
$r_{it}s_i$	.912	.926	.920	1.0	1.0	1.0	1.0
D	.896	.908	.906	2.0	2.0	2.0	2.0
$\phi_{.33}$	.886	.896	.873	3.5	3.0	3.0	3.0
$\phi_{.27}$	.866	.885	.872	3.5	4.0	4.0	4.0
r_{pb}	.857	.864	.859	6.0	7.0	5.0	6.0
r'_{pb}	.839	.865	.850	8.0	6.0	7.0	7.0
S	.850	.843	.844	7.0	12.0	9.5	9.5
$r_{pb,c=.90}$	.794	.868	.844	20.0	5.0	9.5	9.5
r''_{pb}	.837	.853	.842	10.0	8.0	11.5	10.0
$r_{.33t}$	.836	.845	.848	11.0	10.5	8.0	10.5
$\phi_{.10}$	.859	.834	.842	5.0	15.5	11.5	11.5
$r_{pb,c=.50}$	.816	.841	.854	16.0	13.0	6.0	13.0
$r_{pb,c=.75}$	.766	.851	.840	23.0	9.0	13.0	13.0
r'_b	.825	.817	.835	13.0	20.0	14.0	14.0
r_{nb}	.833	.822	.831	12.0	18.0	15.0	15.0
r_b	.817	.813	.819	15.0	21.0	16.0	16.0
$r_{pb,c=.25}$	.807	.845	.800	17.0	10.5	21.0	17.0
r''_b	.822	.809	.813	14.0	22.0	17.0	17.0
$r_{.50t}$	.838	.804	.811	9.0	23.0	18.0	18.0
r_E	.781	.834	.804	21.0	15.5	19.0	19.0
$r_{.27t}$	.801	.824	.791	19.0	17.0	22.0	19.0
$\phi_{.50}$	.804	.820	.790	18.0	19.0	23.0	19.0
r_D	.780	.835	.801	22.0	14.0	20.0	20.0
$r_{.10t}$	.671	.647	.701	24.0	24.0	24.0	24.0

TABLE 4

Rank-Order Correlations Between Population Estimates of the Various Discrimination Indices and Calculated Values of Corresponding Indices Obtained Using Sets of 40 Items

Index	Rho Coefficients (1)	(2)	(3)	Ranks of Rho Coefficients (1)	(2)	(3)	Median Rank
D	.907	.907	.904	1.0	2.0	1.0	1.0
$r_{it^{S}i}$	.902	.910	.902	2.0	1.0	2.0	2.0
$\phi_{.27}$	.882	.871	.864	3.0	3.0	3.0	3.0
$\phi_{.33}$	.864	.864	.857	4.0	4.0	4.0	4.0
$r_{.33t}$	.848	.862	.826	5.5	5.0	13.0	5.5
r_{pb}	.832	.844	.836	10.0	7.0	7.0	7.0
$r_{pb,c=.25}$	.845	.847	.822	7.0	6.0	17.0	7.0
S	.842	.840	.832	8.0	8.0	9.0	8.0
$\phi_{.10}$	.839	.829	.853	9.0	13.0	5.0	9.0
r'_{pb}	.824	.837	.828	11.0	9.0	11.0	11.0
$r_{.50t}$	.848	.828	.827	5.5	14.0	12.0	12.0
$r_{pb,c=.50}$	.810	.831	.839	17.5	12.0	6.0	12.0
$r_{pb,c=.90}$	.823	.822	.835	12.5	15.5	8.0	12.5
r''_{pb}	.823	.835	.824	12.5	10.0	14.0	12.5
r_{nb}	.816	.832	.823	14.0	11.0	15.0	14.0
$r_{.27t}$	.813	.822	.822	15.0	15.5	17.0	15.5
r''_{b}	.804	.818	.822	21.0	18.0	17.0	18.0
r'_{b}	.802	.821	.819	22.5	17.0	19.0	19.0
$r_{pb,c=.75}$	.806	.810	.830	20.0	20.0	10.0	20.0
r_{b}	.802	.815	.811	22.5	19.0	20.0	20.0
r_{F}	.812	.809	.806	16.0	21.5	22.0	21.5
r_{D}	.810	.809	.806	17.5	21.5	22.0	21.5
$\phi_{.50}$	.808	.802	.806	19.0	23.0	22.0	22.0
$r_{.10t}$	.698	.722	.707	24.0	24.0	24.0	24.0

Flanagan's approximation of the product-moment correlation (r_p), and Davis's transformation of the product-moment correlation (r_D) were very similar with respect to their invariance properties. Both coefficients associated with these two indices ranked 15th to 20th relative to other indices. Therefore, r_p and r_D would be classified as being among the more unstable discrimination indices included within the present investigation.

When the point-biserial coefficient was adapted for use with criterion-referenced tests ($r_{pb,c=c_x}$), the resulting statistic demonstrated a degree of invariance slightly less than that of r_{pb}, but slightly more than r_D. Maximum values of the adapted point-biserial coefficient were obtained when the criterion score approached the test mean, the mean for the 50 VR items being 23.375. The stability of $r_{pb,c=c_x}$ appeared to be approximately the same as different criterion scores were used, however there was a slight and perhaps unimportant increase in stability when the criterion score was equal to 50% of the items included in a set; e.g., 50% of 10 items when samples consisting of 10 items were used.

Discussion

Data from the present study indicate that as an item is given membership within varying sets of other items, Gulliksen's $r_{it}s_i$ and Findley's D are the most stable of the 24 discrimination indices investigated. This property held true over several replications within and across sets of 10, 25, and 40 items. The data suggested that when item discrimination is used as a significant factor in determining which items from several forms are selected for inclusion in a final test form, use of $r_{it}s_i$ or D offer more assurance that the discriminative ability of an item as indicated by the index used, will remain relatively constant. For these two indices, the discriminative ability of the items selected is less a function of the chance combinations of items which made up the various trial forms than when the other discrimination indices investigated in the present study are used.

Both the point-biserial and biserial coefficients were modified by the various procedures suggested by Henrysson (1963; 1971). These coefficients both represent correlations between an item and the total test of which the item is a part, and as a result tend to be inflated, particularly for tests involving relatively few items. When these coefficients were calculated using sets of 10 items, the adjustment of these

correlations (r'_{pb} and r'_b) suggested by Henrysson (1963) for the removal of this spuriousness not only resulted in a lower discrimination coefficient, but also in a less stable coefficient. This decrease in stability can be attributed to the fact that as an item is given membership within different sets of items, the removal of the particular item of concern from each composite of items results in a reduction of the commonality associated with each of the adjusted composite measures. Irregardless, if adjustment is or is not made for spuriousness, the differences in invariance properties decrease for the biserial and point-biserial coefficients when computation of these indices are based on sets of 25 and 40 items. The existence of this characteristic suggests that the indicated spuriousness, as it relates to the invariance property of the biserial and point-biserial coefficients, is insignificant with the possible exception of small sets of items. Henrysson further reasoned that coefficients such as r'_{pb} and r'_b would result in items being correlated with different "remainders of items" and likewise slightly different contents. The readjusted coefficients (r''_{pb} and r''_b) supposedly express the correlation between the various items and the general factor being measured by the test, thus eliminating this problem of different sums. If this were the case, one would expect r''_{pb} to be more stable than r'_{pb} and r''_b more stable than r'_b. However, data from the present study does not consistently substantiate such relationships.

Flanagan's approximation of the product-moment correlation coefficient was one of the less stable indices included in this investigation. Flanagan (1939) supported the use of the biserial coefficient as an item discrimination index, and proposed r_D as a"short-cut" procedure for estimating r_b. For this purpose, an abac was constructed which provided estimates of the biserial coefficient on the basis of the proportion of subjects in the upper and lower criterion groups which correctly answered the item. However, with the present availability of computers, and the relative instability of this index, the advantage of such an estimation procedure is minimal in most situations.

Davis' transformation of the product-moment correlation coefficient was basically a Fisher's Z transformation of biserial correlation coefficients. Its intent was to provide an index with which arithmetic processes could be carried out. Davis' transformation was originally developed for use with Flanagan's method of estimating the biserial coefficient. In fact, David (1946) developed a chart which would estimate the transformed index directly from the performances of subjects in upper and lower criterion groups. Since the Davis transformation is quadratical, the rank-ordering of index values before and after transformation would be identical, thus explaining the similar invariance properties of r_F and r_D. The Davis transformation

could also appropriately be used with r_b, and for reasons just stated, would then exhibit invariance properties similar to the biserial coefficient.

Estimates of r_t used in this study were based on tables prepared by Davidoff and Goheen (1953), these tables in turn representing solutions of the cosine-pi formula for approximating r_t. It may be that the apparent instability of r_t reported in the present study is a function of this estimation procedure instead of a characteristic of the tetrachoric coefficient itself. However, Bonvier, Perry, and Hertzka (1954) demonstrated close approximation of r_t is obtained when the cosine-pi formula is used as the method of estimating the tetrachoric coefficient as long as both variables are dichotomized at their medians. Since such a condition was prevalent in the present study, it is questionable whether the considerable computational labor involved in the direct computation of the tetrachoric coefficient would be rewarded by a more stable index of discrimination.

If discrimination indices are relevant to criterion-referenced tests for the purpose of selecting items from item pools, then the invariance of these indices as an item is moved from one test form to another is of significant concern. The indices discussed above are appropriate for use with norm-referenced tests, however using procedures discussed by Livingston (1972), the present study proposed an adaptation of the point-biserial coefficient which may be appropriate for use with a criterion-referenced test. The point-biserial coefficient was selected on the basis of its present popularity as a discrimination index relative to norm-referenced tests, and also on the basis of its ease of conversion using the techniques suggested by Livingston.

The data obtained from the present study indicate that the proposed index has approximately the same degree of invariance as does the point-biserial or biserial coefficient. When various criterion scores were used, indices were slightly more stable when the criterion score selected was closest to the actual test mean. This relationship however, was not consistent, therefore conclusions based on the data would be somewhat speculative.

The only difference between formulas associated with the traditional forms of the point-biserial coefficient and the point-biserial adapted for use with criterion-referenced tests is that the former uses the standard deviation of test scores whereas the latter uses the square root of what Livingston refers to as the mean squared deviation of scores from the criterion score. As with most of Livingston's modifications of classical test theory, the indicated deviation of scores is dependent on each scores distance from criterion score, in contrast to differences from the mean score. Because the mean squared deviation of scores from the criterion score takes on a minimal value when the criterion score is equal to the test mean, the

variance represents the minimal value of Livingston's mean squared deviation of scores. Similarly, because either the standard deviation or square root of the mean deviation of test scores appears in the denominator of the equation for computing the point-biserial coefficient, as the mean deviation of scores increases (i.e., as the criterion score deviates further from the test mean), the value of the point-biserial coefficient adjusted for criterion tests decreases. As a result, it is very important that if the adapted index is used as a basis of selecting items for inclusion in a criterion-referenced test, the test constructor must realize that such an index will be inflated whenever the mean test score approaches the criterion score, and also that this discrimination index should be calculated assuming the same criterion score used in the original analysis is incorporated in the final form of the test being constructed.

The phi coefficient would intuitively represent an appropriate index of item discrimination for use with criterion-referenced tests. Subject performance above or below the criterion score would represent one dichotomy, whereas successful or unsuccessful performance with a particular item would represent the other dichotomy. Although supporting empirical evidence is not known to exist, it is inferred from data obtained in the present study that the phi coefficient would be relatively unstable as a discrimination index. The phi coefficient was considerably less stable when computations were based on Ss scoring in the upper and lower 50% than when based on Ss scoring in the upper and lower 33% or upper and lower 27%. It is expected that if Ss were dichotomized at some point other than the median, such as would be the case if Ss were dichotomized on the basis of their performance relative to a criterion score, the resulting phi coefficient would be less stable than had only Ss scoring at the extremes of the continuum been used.

Brennan (1972) suggested a generalized form of Findley's D which allows for deviant numbers of subjects to be included in the upper and lower criterion groups. Brennan indicated that this particular index would be appropriate for use with criterion-referenced tests if the upper and lower criterion groups were determined on the basis of Ss' relative performance on the criterion score. However, data from the present study suggest that when determination of an index required subjects to be dichotomized on the basis of their performance relative to a specific point, the resulting index is relatively unstable.

The indices which demonstrated the greatest degree of invariance tended to be those which were dependent in part on the difficulty of the item with which they were associated. Since difficulty for each item was not allowed to vary for respective items, it is possible that an index's dependence on item difficulty acted as a stabilizing agent. Authorities, such as Nunnally (1967) and Henrysson (1971) disagree relative to the effect difficulty should have on the discrimination index. If the view were taken that item discrimination should be independent

of item difficulty, the [illegible] coefficient adjusted by the maximum point-biserial would represent the index with the greatest degree of invariance.

The general finding of the present investigation was that the relative discriminative of a particular item, irregardless of which index was used to assess this discrimination, was highly invariant when items were placed within the context of varying sets of items. The most notable exceptions of this were estimates of the discriminative based on the responses of Ss scoring in the upper and lower 10%. Some consistent differences between indices, relative to the invariance property investigated in this study, were observed. For example, Gulliksen's item reliability index and Findley's difference index were consistently the least variant of the indices included in the investigation.

References

Bouvier, E., Perry, N., Michael, W., and Hertzka, A. A study of the error in the cosine-pi approximation to the tetrachoric coefficient of correlation. Educational and Psychological Measurement, 1954, 14, 690-693.

Brennan, Robert L. A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

Brogden, H. E. A new coefficient: application to biserial correlation and to estimation of selective efficiency. Psychometrika, 1949, 14, 169-182.

Davidoff, M.D. and Goheen, H. W. A table for the rapid determination of the tetrachoric correlation coefficient. Psychometrika, 1953, 18, 115-121.

Davis, F. B. Item analysis data, their computation, interpretation, and use in test construction. Harvard Educational Papers, 1946, No. 2.

Findley, W. G. A rationale for evaluation of item discrimination statistics. Educational and Psychological Measurement, 1956, 16, 175-180.

Flanagan, J. C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. Journal of Educational Psychology, 1939, 30, 674-680.

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, Inc., 1950.

Henrysson, S. Correction of item-total correlations in item analysis. Psychometrika, 1963, 28, 211-218.

Henrysson, S. Gathering, analyzing and using data on test items. In R. L. Thorndike (Ed.), Educational Measurement. (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Livingston, S.A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 2, 13-29.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Oosterhof, A. C. Invariance properties of item discrimination indices as a function of item inclusion within different item sets. Unpublished doctoral dissertation, University of Kansas, 1972.